Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Information Theoretical Aspects of Complex Systems

Lecture 2.02

EEU45C09 / EEP55C09
Self Organising Technological Networks

CONNECT
Networks of the Future

Nicola Marchetti
nicola.marchetti@tcd.ie

Measuring Complexity

- In practice, we study the systems that interest us, for whatever reasons

- Having chosen a system to study, we might well ask: *How complex is this system?*

- In general, we want at least to be able to *compare* two systems, and be able to say that a certain system *A* is more complex than another system *B*

- Eventually, we would like to have some sort of *numerical rating* scale

Measuring Complexity (2)

- *Various approaches* to the above task have been proposed, among them:

1. Human observation and (subjective) rating

2. Number of parts or distinct elements (what counts as a distinct part?)

3. Dimension (measured how?)

4. Number of parameters controlling the system

Measuring Complexity (3)

- *Various approaches* to the above task have been proposed, among them:

5. Minimal description (in which language?)

6. Information content (how do we define/measure information?)

7. Minimal generator/constructor (what machines/methods can we use?)

8. Minimum energy/time to construct (how would the evolution of the system count?)

Measuring Complexity (4)

- Most (if not all) of these measures will actually be measures associated with a *model* of a phenomenon

- Two observers (of the same phenomenon?) may develop or use very different models, and thus disagree in their assessments of the complexity

- For example, counting the number of parts is likely to depend on the *scale* at which the phenomenon is viewed (e.g., counting atoms is different from counting molecules, cells, organs, etc.)

Measuring Complexity (5)

- We shouldn't expect to be able to come up with a single universal measure of complexity

- The best we are likely to have is a measuring system useful to a particular observer, in a particular context, for a particular purpose

- Our focus will be on measures related to how *surprising or unexpected* an observation or event is

- We call this approach *information theory*

Basics of Information Theory

- We would like to develop a *usable measure of the information* we get from observing the occurrence of an event having *probability p*

- Our first reduction will be to ignore any particular features of the event, and only observe whether or not it happened

- Thus we will think of an event as the observance of a symbol whose probability of occurring is p

- We will thus be defining the information in terms of the probability p

Basics of Information Theory (2)

- The approach we will be taking is *axiomatic*: we will see in a while a list of the four fundamental axioms we will use

- Note that we can apply this axiomatic system in any context in which we have available a set of non-negative real numbers

- A specific special case of interest is *probabilities* (i.e., real numbers between 0 and 1)

Basics of Information Theory (3)

- We want our information measure $I(p)$ to have the following *axioms*:

1. Information is a non-negative quantity: $I(p) \geq 0$

2. If an event has probability 1, we get no information from the occurrence of the event: $I(1) = 0$

Basics of Information Theory (4)

3. If two *independent* events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two information: $I(p_1 \cdot p_2) = I(p_1)+I(p_2)$

4. The information measure is a continuous (and monotonic) function of the probability (slight changes in probability should result in slight changes in information)

Basics of Information Theory (5)

❑ From the axioms we can derive the following *properties* (assuming independent events, same *p*):

1. $I(p^2) = I(p \cdot p) = I(p) + I(p) = 2 \cdot I(p)$

2. Thus in general, $I(p^n) = n \cdot I(p)$

3. $I(p) = I\left(\left(p^{1/m}\right)^m\right) = mI\left(p^{1/m}\right) \Rightarrow I\left(p^{1/m}\right) = \frac{1}{m} I(p)$

4. Thus in general, $I\left(p^{n/m}\right) = \frac{n}{m} \cdot I(p)$

5. By continuity, for $0<p\leq 1$, and $a>0$ real number

$$I(p^a) = a \cdot I(p)$$

Basics of Information Theory (6)

- *I(p)*, as it is defined by the axioms and consequent properties, for $0<p\leq 1$, can be identified with the *logarithm*

$$I(p) = -\log_b(p) = \log_b(1/p)$$

for some base $b>0$.

Basics of Information Theory (7)

- Summarising, from the four axioms

1. $I(p) \geq 0$
2. $I(1) = 0$
3. $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$
4. *I(p)* is monotonic and continuous in *p*

we can derive that

$$I(p) = -\log_b(p) = \log_b(1/p)$$

for some positive constant *b*.

- The base *b* determines the units we are using

Basics of Information Theory (8)

- We can change the units by changing the base
- Indeed, for $b_1, b_2, x > 0$

$$x = {b_1}^{\log_{b_1}(x)}$$

- Therefore,

$$\begin{aligned}\log_{b_2}(x) &= \log_{b_2}\left({b_1}^{\log_{b_1}(x)}\right) = \\ &= \left(\log_{b_2}(b_1)\right) \cdot \left(\log_{b_1}(x)\right)\end{aligned}$$

Basics of Information Theory (9)

- Thus, using different bases for the logarithm results in information measures which are just constant multiples of each other, corresponding with measurements in different units:

  - ✓ log_2 units ➔ *bits* (from 'binary')
  - ✓ log_3 units ➔ *trits* (from 'trinary')
  - ✓ log_e units ➔ *nats* (from 'natural' logarithm; we can also use the *ln* notation)
  - ✓ log_{10} units ➔ *Hartleys* (after an early worker in the field)

Example

- Flipping a fair coin once will give us events *h* and *t* each with probability *1/2*, and thus a single flip of a coin gives us $-\log_2(1/2) = 1$ *bit of information* (whether it comes up *h* or *t*)

- Flipping a fair coin n times (or, equivalently, flipping n fair coins) gives us $-\log_2((1/2)^n) = \log_2(2^n) = n\log_2(2) = n$ bits of information

- We could enumerate a sequence of *5* flips as, for example: *hthht* or, using *1* for *h* and *0* for *t*, the *5* bits *10110*

- We thus get the nice fact that *n* flips of a fair coin gives us *n* bits of *information*, and takes *n* *binary digits* to specify. That these two are the same reassures us that we chose a *good definition* of our information measure

Entropy Theory

- Suppose now that we have n symbols $\{a_1, a_2, \ldots, a_n\}$, and some source is providing us with a stream of these symbols

- Suppose further that the source emits the symbols with probabilities $\{p_1, p_2, \ldots, p_n\}$, respectively

- We also assume that the symbols are emitted *independently* (successive symbols do not depend in any way on past symbols)

- What is the *average amount of information* we get from each symbol we see in the stream?

Entropy Theory (2)

- What we want here is a *weighted average*

- If we observe the symbol a_i, we will be getting $log(1/p_i)$ information from that particular observation

- In a long run (say N) of observations, we will see (approximately) $N \cdot p_i$ occurrences of symbol a_i

- Thus, in the N (independent) observations, we will get total information I of

$$I = \sum_{i=1}^{n} (N \cdot p_i) \cdot \log(1/p_i)$$

Entropy Theory (3)

- But then, the average information we get per symbol observed will be

$$\frac{I}{N} = \frac{1}{N}\sum_{i=1}^{n}(N \cdot p_i) \cdot \log(1/p_i) = \sum_{i=1}^{n} p_i \cdot \log(1/p_i)$$

- Note that, since $\lim_{x \to 0} x \log(1/x) = 0$, we can, for our purposes, define $p_i \log(1/p_i) = 0$ when $p_i = 0$

Entropy Theory (4)

- This brings us to a fundamental definition

- This definition is due to *Shannon* in 1948, in the seminal papers in the field of information theory

- We have defined information strictly in terms of the probabilities of events. Therefore, let us suppose that we have a set of probabilities (a probability distribution) $P = \{p_1, p_2, \ldots, p_n\}$

- We define the *entropy* of the distribution P by

Strictly speaking, this is a discrete density function – but is it common to use density/distribution interchangeably (context tells us)

$$S(P) = \sum_{i=1}^{n} p_i \cdot \log(1/p_i)$$

Entropy Theory (5)

- The generalisation to a *continuous* probability distribution is

$$S(P) = \int P(x) \cdot \log(1/P(x))\, dx$$

- Another way to think about entropy is in terms of *expected value.*

- *Let us consider* a discrete probability distribution $P=\{p_1, p_2, \ldots, p_n\}$, with $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$

, or a continuous distribution $P(x)$ with

$P(x) \geq 0$ and $\int P(x) dx = 1$

Entropy Theory (6)

- We can define the expected value of a discrete set $F=\{f_1,f_2,\dots,f_n\}$ (associated with the discrete distribution P), or of a function $F(x)$ (associated with the continuous distribution $P(x)$) by

$$< F >= \sum_{i=1}^{n} f_i p_i$$

or

$$< F(x) >= \int F(x)P(x)dx$$

Entropy Theory (7)

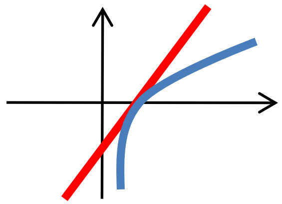
- With the above definitions we have that

$$S(P) = \langle I(P) \rangle$$

- In other words, *the entropy of a probability distribution is the expected value of the information of the distribution*

The bridge between information theoretical and physical entropies lies in *how hard it is to describe a physical state or process*. The amount of information it takes to describe something is proportional to its entropy.

Gibbs Inequality



- The tangent to *ln(x)* at *x=1* is the line *y=x-1*. Further, since *ln(x)* is concave down, we have that, for *x>0*, *ln(x)* $\le$ *x-1* (with equality only when *x=1*)

- Now, given two probability distributions $P=\{p_1,p_2,\ldots,p_n\}$, $Q=\{q_1,q_2,\ldots,q_n\}$, where $p_i, q_i \ge 0$ and

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$$, we have

$$\sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right) \underset{\ln x \le x-1}{\le} \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_{i=1}^{n} (q_i - p_i) = \qquad (1)$$

$$= \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0$$

Gibbs Inequality (2)

- The equality in (1) holds only when $p_i=q_i$ for all i
- Gibbs Inequality holds for any base of the logarithm, not just e
- We can use Gibbs Inequality to find the probability distribution which *maximises* the entropy function
- Suppose $P=\{p_1,p_2,\ldots,p_n\}$ is a probability distribution

Maximum entropy

$$S(P) - \log(n) = \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) =$$

$$= \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) \sum_{i=1}^{n} p_i =$$

$$= \sum_{i=1}^{n} p_i \log(1/p_i) - \sum_{i=1}^{n} p_i \log(n) =$$

$$= \sum_{i=1}^{n} p_i \left[\log(1/p_i) - \log(n)\right] =$$

$$= \sum_{i=1}^{n} p_i \left[-\log(p_i) + \log(1/n)\right] =$$

$$= \sum_{i=1}^{n} p_i \log\left(\frac{1/n}{p_i}\right) \underset{Gibbs}{\leq} 0$$

Equality holds only when $p_i = 1/n$ for all i

Minimum and maximum entropy

- The above calculation, and the fact that the entropy is the expected value of an information (and thus non-negative, by the first axiom of information) imply that

$$0 \leq S(P) \leq \log(n)$$

- $S(P)=0$ when exactly one of the p_i's is *1* and all the rest are *0*

- $S(P)=\log(n)$ only when all of the events have the same probability $1/n$, i.e., *the maximum of the entropy function is the log() of the number of possible events, and occurs when all the events are equally likely*

Example

- How much information can a student get from a single grade?

- First, the maximum information occurs if all grades have equal probability (e.g., in a pass/fail class, on average half should pass if we want to maximize the information given by the grade)

- The maximum information the student gets from a grade will be:

  - ✓ Pass/Fail : 1 bit ($log_2(2)=1$)
  - ✓ A, B, C, D, E : 2.3 bits ($log_2(5)=2.3$)

- Thus, using five grades instead of just pass/fail gives the students about 1.3 more bits of information

Remarks on Information and Entropy

- These definitions of information and entropy may not match with some *other uses of the terms*

- For example, if we know that a source will, with equal probability, transmit either the complete text of Hamlet or the complete text of Macbeth (and nothing else), then receiving the complete text of Hamlet provides us with precisely 1 bit of information (and *nothing more than that*)

- Suppose a book contains ascii characters. If the book is to provide us with information at the maximum rate, then each ascii character will occur with equal probability (it will then be a *random sequence* of characters)

Remarks on Information and Entropy (2)

- It is important to recognize that our definitions of information and entropy *depend only* on the probability distribution

- In general, it *won't make sense* for us to talk about the information or the entropy of a source without specifying the probability distribution

- Beyond that, it can certainly happen that two different *observers* of the same data stream have different *models* of the source, and thus associate different probability distributions to the source

- The two observers will then assign different values to the *information and entropy* associated with the source

Remarks on Information and Entropy (3)

- This observation to certain extent accords with our *intuition*

- For example, two people listening to the same lecture can get very different information from the lecture

- Without appropriate background, one person might not understand anything at all, and therefore have as probability model a *completely random* source, and therefore get much more information than the listener who understands quite a bit, and can therefore anticipate much of what goes on, and therefore assigns *non-equal probabilities* to successive words